Supplementary Information

1 Optimization algorithm

2 Results on HIV data

We incorporate an additional task: HIV (human immunodeficiency virus), which is quite dissimilar to our original setting involving three viruses (HIV is a retrovirus). This experiment tests how our multitask model behaves in the presence of a relatively outlier task. The interactions data has ≈4000 human-HIV PPI

Algorithm 1: Alternating Least Squares

```
Input:
  $k$ : number of latent factors
  $\Gamma$: pairs of entities for initialization
  For every task $t$,
  $\{\mathbf{x}_{ti}\}, \{\mathbf{y}_{tj}\}$: feature vectors for pathogen and host proteins resp.
  $\Omega_t$: the observed entries of the matrix $M_t$
Initialization:
An iteration $r$, let $\mathbf{S}^r$ represent $\{S_t^r\}_{t=1}^T$
$\mathbf{S}^0 \leftarrow 0$
$U^0 \leftarrow$ top-$k$ left singular vectors and $V^0 \leftarrow$ top-$k$ right singular vectors from the SVD of $\sum_{(i,j)\in\Gamma} \mathbf{x}_i \mathbf{y}_j^\intercal$
$\mathcal{L}^0$ : initial loss
repeat
    $U^{r+1} \leftarrow \underset{U}{\operatorname{argmin}}\, \mathcal{L}(U, V^r, \mathbf{S}^r)$
    $V^{r+1} \leftarrow \underset{V}{\operatorname{argmin}}\, \mathcal{L}(U^{r+1}, V, \mathbf{S}^r)$
    For each task $t$
        $S_t^{r+1} \leftarrow \underset{S_t}{\operatorname{argmin}}\, \mathcal{L}(U^{r+1}, V^{r+1}, \mathbf{S}_{-t}^r)$
    Compute $\mathcal{L}^{r+1}$ and let $\delta \leftarrow (\mathcal{L}^r - \mathcal{L}^{r+1})/\mathcal{L}^r$
until $\delta < \tau$
```

across several strains of which 1320 are non-homologous PPI. Please refer to the supplementary for the results.

This experiment particularly tests the robustness of all the multitask learning approaches by the addition of an unrelated task. Table 1 shows the results.

	10 % training			
	Ebola	*Hep-C*	*Influenza*	*HIV*
Homolog	0.230±.06	0.178±.01	0.158±.01	0.170 ±.01
STL	0.189±.09	0.702±.08	0.286±.02	0.208 ±.01
Sparse+LR	0.135 ±.06	0.724 ±.05	0.280 ±.01	0.205 ±.03
BSL-MTL	**0.241** ±.01	**0.793** ±.06	**0.459** ±.01	**0.354** ±.01

Table 1: AUC-PR upon addition of a new task - HIV, in the 10% setting. Note that STL and the homolog method build independent models so their AUC numbers are the same as in Table 2. Among the multitask learning methods, BSL-MTL's performance is not hurt by the addition of an unrelated task such as HIV.

3 Parameter tuning

We tune the hyper parameters using a 3 fold cross-validation (CV) on the training split. For all baseline regularization parameters we tried the range: $[100, 50, 10, 1, 0.1, 0.01, 0.05, 10^{-3}]$. To address the class-skew we assign a higher weight to the positives. For BSL-MTL, to tune the rank parameter 'k' we tried: [1, 5, 10, 25, 40, 60, 100] and the regularization parameter controlling the norm of U and V was tuned over the range λ=$\{1...10^{-3}\}$. For each task t, σ_t was varied over the values $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ and μ_t was varied over $\{0, 0.25, 0.5, 0.75, 1\}$. The optimal setting was: $k = 10$.

4 Biological significance of interactions

We analyze sequence motifs derived from the top k-mers that contribute to interactions. The significant entries of the model parameters U, V and $\{S_t\}$ were used to compute these motifs. The top positive-valued entries from the product UV^T indicate which pairs of features: $((f_v, f_h)$: virus protein feature, human protein feature) are important for interactions across all the virus-human PPI tasks. Analogously, the entries from S_t give us pairs of features important to a particular virus-human task 't'. We find that most of the top entries from UV^T correspond to linear virus features, whereas those from the various S_t involve bilinear features.

We analyze the k-mers corresponding to the top 20 features from each of the matrices.

Note that our features do not directly correspond to a unique amino-acid k-mer (see Section 4.2): the virus feature f_v will map to several amino-acid sequences (for instance KKCC, KRCC, RRCC etc all map to a single feature due to the molecular similarity between the amino acids K and R being both positively charged). Given the set of top virus features we can obtain the corresponding set of amino-acid k-mers, say AA_v, by reversing the feature-generation step. However most of the possible k-mers do not appear in the training data (ex: out of the 160,000 (=20^4) possible 4-mers ≈24,000 appear). Let AA_{tr} be the set of amino-acid k-mers that appear in the training data. Then, the intersection $I_v = AA_v \cap AA_{tr}$ gives us the important amino-acid k-mers from virus proteins w.r.t interaction prediction. To summarize I_v, we use a popular tool Seq2Logo to generate a sequence motif. The logos for the two-, three-, four-mers from I_v are generated independently. Since we only want to summarize, we use the Shannon logo type (which does not consider any background amino-acid distribution) with the following settings: clustering-method=None, weight-on-prior=1 (pseudo-counts do not make sense in our analysis). Figure 1 shows the motif that is common across viruses. We observe that the shared tri-mer motif for virus proteins in Figure 1 is dominated by hydrophilic amino acids (T, K, R, D, E).

This procedure described above is used to analyze the most significant human protein features, obtained from the matrix UV^T. These motifs are shown in the supplementary. The task-specific features i.e significant k-mers from virus-proteins of *Ebola*, *Hepatitis* and *Influenza* are obtained from the matrices S_{ebola}, S_{hepc} and S_{flu} respectively. The motif for Hepatitis is shown in Figure 1(right); the rest can be found in the supplementary. The virus-specific motifs seem to be dominated by hydrophobic residues (I, P, L, V, A, G) though S and T do appear in some motifs as well.

Phosphorylation sites: We found the frequent occurrence of S and T and sometimes Y in the motifs striking and suspected this may be related to the amino acids being frequent targets of phosphorylation. Phosphorylated sites often serve as PPI sites, and databases such as Phosphosite are repositories for known sites in human proteins. Since these are sites in human proteins, we searched for the patterns from the 4-mer motif in Fig 8 and found several to be flanking known phosphorylation sites in human proteins: `LLLs`, `LLLt`, `ILLs`, `PPPs`, `PIPs`, `PIPt`, `LIPs`, `PLLt` (lower-case indicates the putative phosphorylation site). This observation also supports the notion that the motifs predictive of interaction are biologically significant.

Evidence in IEDB[1]: We found experimental evidence for the significance for the virus motifs in the Immune Epitope[2] database (IEDB). The pattern `IVGG` from the *Hepatitis-C* motif in Fig. 6 is found in 53 epitopes. From

[1] `www.iedb.org`

[2] An epitope is a very short sequence from the virus that binds to human antibodies

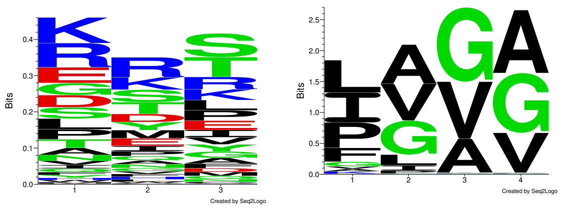


Figure 1: (*Left*) Top tri-mer sequence motifs from virus proteins across all three viruses. These correspond to features important for interactions and is derived from our model parameter $H = UV^T$. (*Right*) Motif corresponding to enriched 4-mer patterns found by our model (parameter: S_{hepc}) for the Hepatitis-C virus task.

the *Ebola* motifs in Fig 2, we find that `TLAT` is part of six different epitopes, `SLTT` appears in three epitopes. `PLIK`, `SLLL` from the *Influenza* motif are also found in many epitopes. Finding that the virus k-mer patterns predicted by our method are recognized by human antibodies is a further validation of its performance. Further, using higher dimensional k-mers (where k=7, 8, 9) as features in our model will give motifs from which complete epitopes can be derived. Our model thus has applications in epitope prediction as well, where conventional methods consist of scanning all possible k-mers from protein sequences to identify likely epitopes.

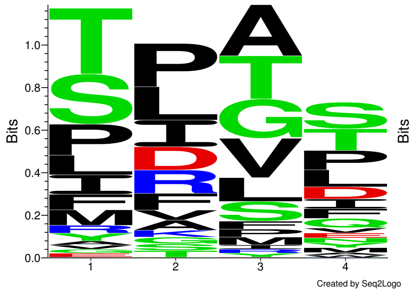


Figure 2: Sequence motif from top four-mer features specific to *Ebola* proteins

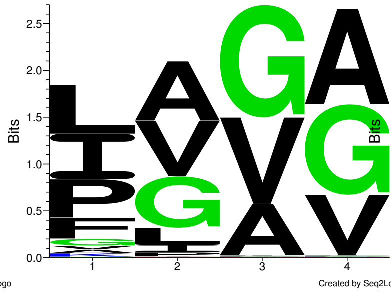


Figure 3: Sequence motifs specific to *Hepatitic-C* proteins that are also important to interactions.

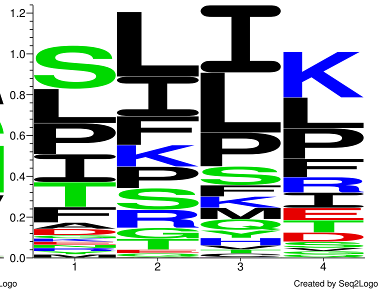


Figure 4: Sequence motifs specific to *Influenza* proteins that are also important to interactions.

5 Tri-mers

Below, we show the sequence motifs from the tri-mers found to be highly relevant to predicting interactions between human and viral proteins.

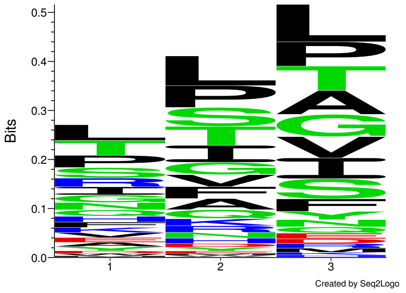


Figure 5: Sequence motifs specific to *Ebola* proteins that are also important to interactions.

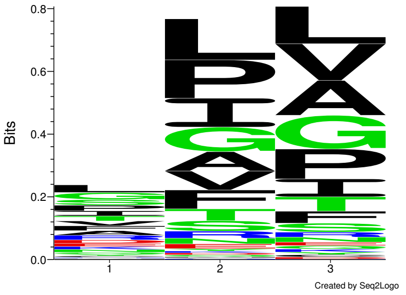
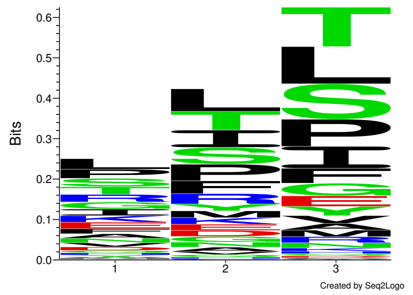


Figure 6: Sequence motifs specific to *Hepatitic-C* proteins that are also important to interactions.

References

[1] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham,



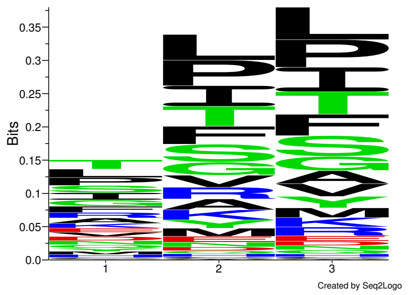


Figure 7: Sequence motifs specific to *Influenza* proteins that are also important to interactions.

Figure 8: Sequence motif constructed from the top tri-mer features of human proteins

and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2015.

[2] Randi Vita, James A Overton, Jason A Greenbaum, et al. The immune epitope database (iedb) 3.0. *Nucleic acids research*, 43(D1):D405–D412, 2015.